Amaya Nogales Gómez

Emilio Carrizosa · Dolores Romero Morales

Mathematical and Algorithmic Sciences Lab, Huawei France R&D, Paris, France

Clustering Categories in Support Vector Machines

PGMO DAYS 2016

November 9, 2016

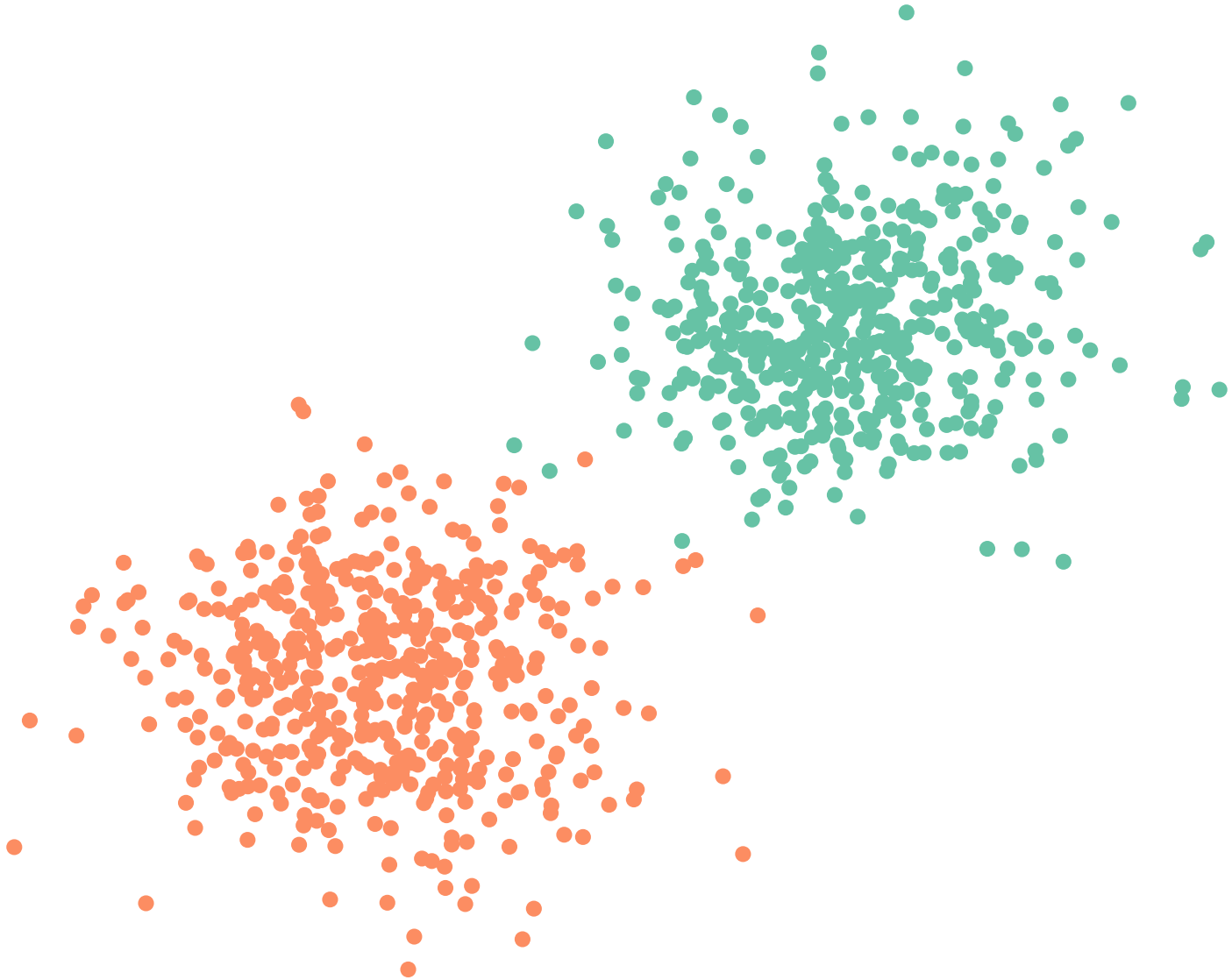
Support Vector Machines (SVM)

- Ω: the population
- Population is partitioned into two classes, $\{-1, +1\}$
- For each object i in Ω, we have
  - $x_i' \in X' \subset \mathbb{R}^{J'}$: vector of continuous features
  - $y_i \in \{-1, +1\}$: class membership

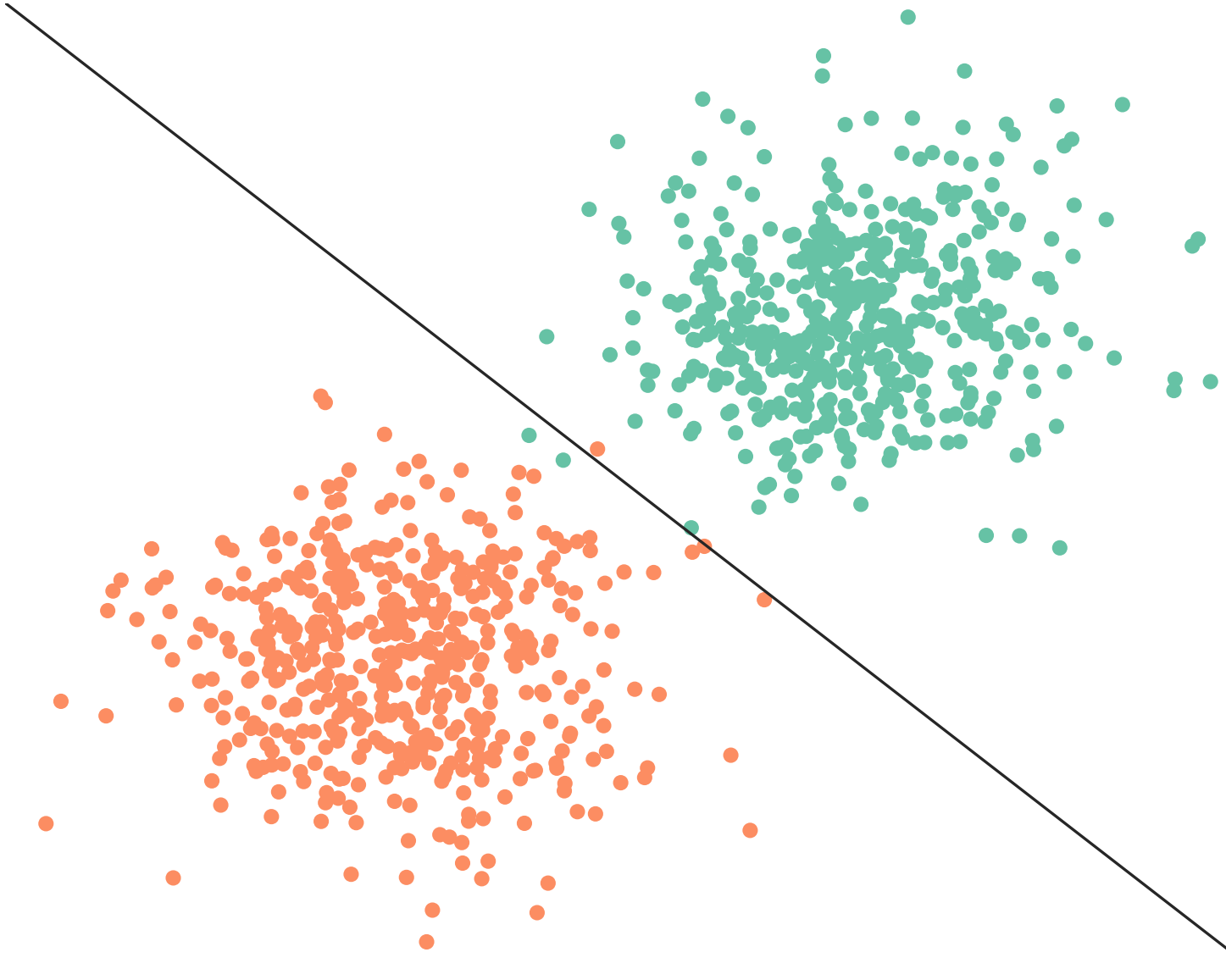
- The goal is to find a hyperplane $(\omega')^\top x' + b = 0$ that aims at separating, if possible, the two classes
- Future objects will be classified as

$$y_i = +1 \quad \text{if} \quad (\omega')^\top x' + b > 0$$
$$y_i = -1 \quad \text{if} \quad (\omega')^\top x' + b < 0$$

The SVM



The SVM



Presence of categorical features

- J categorical features, where j has K_j categories
- Categories are split into 0-1 dummy features:
  - $x_{i,j,k}$: 1 if the value of categorical feature j in object i is equal to category k, 0 otherwise
  - $x_i = (x_{i,j,k}) \in X \subset \{0,1\}^{\sum_{j=1}^{J} K_j}$: vector of categorical features

  Hyperplane:
  $(\omega')^\top x' + b = 0 \rightarrow (\omega)^\top x + (\omega')^\top x' + b = 0$

The SVM formulation

$$\min_{\omega,\omega',b,\xi} \sum_{j=1}^{J}\sum_{k=1}^{K_j} \frac{(\omega_{j,k})^2}{2} + \sum_{j'=1}^{J'} \frac{(\omega'_{j'})^2}{2} + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

s.t. (SVM)

$$y_i\Big(\sum_{j=1}^{J}\sum_{k=1}^{K_j} \omega_{j,k}\, x_{i,j,k} + (\omega')^\top x'_i + b\Big) \geq 1 - \xi_i \qquad \forall i = 1,\dots,n$$

$$\xi_i \geq 0 \qquad \forall i = 1,\dots,n$$

$$\omega \in \mathbb{R}^{\sum_{j=1}^{J} K_j}$$

$$\omega' \in \mathbb{R}^{J'}$$

$$b \in \mathbb{R},$$

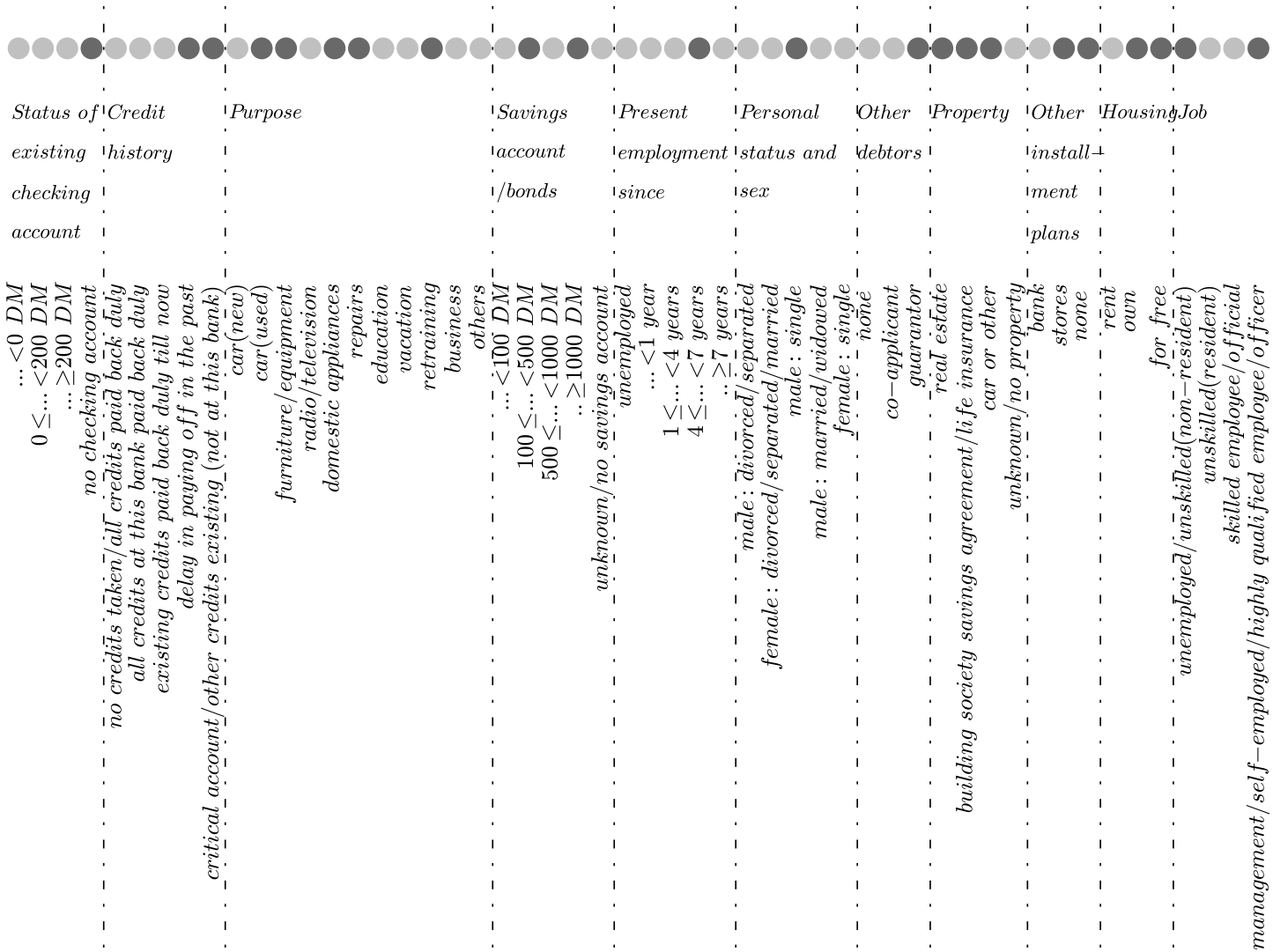
where n is the size of the sample and $\sum_{i=1}^{n} \xi_i$ the most popular choice for the loss function.

Cluster Support Vector Machines (CLSVM)

We propose:

- A methodology to reduce the complexity of the SVM classifier, the Cluster Support Vector Machines (CLSVM)
- Clustering the K_j categories of categorical feature j into L_j clusters, $\forall j$
- Obtaining an SVM-type classifier with categories clustered around their peers
- Without compromising accuracy

An example for $L_j = 2$



The CLSVM methodology

Given a dataset Ω:

1. Find the assignment vector $z^* = (z^*_{j,k,\ell})$, defining a clustering for the categorical features

2. Obtain the clustered dataset $\bar{\Omega}$
   $(y_i, x_i, x'_i) \rightarrow (y_i, \bar{x}_i, x'_i)$
   where $\bar{x}_i = (\bar{x}_{i,j,\ell})$ and $\bar{x}_{i,j,\ell} = \sum_{k=1}^{K_j} z^*_{j,k,\ell} x_{i,j,k}$

3. Find the *CLSVM* classifier for $\bar{\Omega}$, $(\bar{\omega})^\top \bar{x} + (\omega')^\top x' + b = 0$

A Mixed Integer Nonlinear Programming (MINLP) formulation

$$\min_{\bar{\omega},\omega',b,\xi,z} \sum_{j=1}^{J} \sum_{\ell=1}^{L_j} \frac{(\bar{\omega}_{j,\ell})^2}{2} + \sum_{j'=1}^{J'} \frac{(\omega'_{j'})^2}{2} + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

s.t. (CL)

$$y_i \left(\sum_{j=1}^{J} \sum_{\ell=1}^{L_j} \bar{\omega}_{j,\ell} \sum_{k=1}^{K_j} z_{j,k,\ell}\, x_{i,j,k} + (\omega')^\top x'_i + b \right) \geq 1 - \xi_i \qquad i = 1, \ldots, n$$

$$\sum_{\ell=1}^{L_j} z_{j,k,\ell} = 1 \qquad j = 1, \ldots, J; k = 1, \ldots, K_j$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, n$$

$$z \in \{0,1\}^{\sum_{j=1}^{J} L_j K_j}$$

$$\bar{\omega} \in \mathbb{R}^{\sum_{j=1}^{J} L_j}$$

$$\omega' \in \mathbb{R}^{J'}$$

$$b \in \mathbb{R}$$

An MINLP formulation

$$\min_{\bar{\omega},\omega',b,\xi,z} \sum_{j=1}^{J}\sum_{\ell=1}^{L_j} \frac{(\bar{\omega}_{j,\ell})^2}{2} + \sum_{j'=1}^{J'} \frac{(\omega'_{j'})^2}{2} + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

s.t. (CL)

$$y_i \left(\sum_{j=1}^{J}\sum_{\ell=1}^{L_j} \bar{\omega}_{j,\ell} \sum_{k=1}^{K_j} z_{j,k,\ell}\, x_{i,j,k} + (\omega')^\top x'_i + b \right) \geq 1 - \xi_i \qquad i = 1, \ldots, n$$

An MINLP formulation

$$\sum_{\ell=1}^{L_j} z_{j,k,\ell} = 1 \qquad j = 1, \ldots, J; k = 1, \ldots, K_j$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, n$$

$$z \in \{0,1\}^{\sum_{j=1}^{J} L_j \, K_j}$$

$$\bar{\omega} \in \mathbb{R}^{\sum_{j=1}^{J} L_j}$$

$$\omega' \in \mathbb{R}^{J'}$$

$$b \in \mathbb{R}$$

Theoretical results

Proposition

For any optimal solution of CL, given a categorical feature j^*, if there exists ℓ^* such that $z_{j^*,k,\ell^*} = 1 \; \forall k = 1, \ldots, K_{j^*}$, then $\bar{\omega}_{j^*,\ell} = 0 \; \forall \ell = 1, \ldots, L_{j^*}$.

Corollary

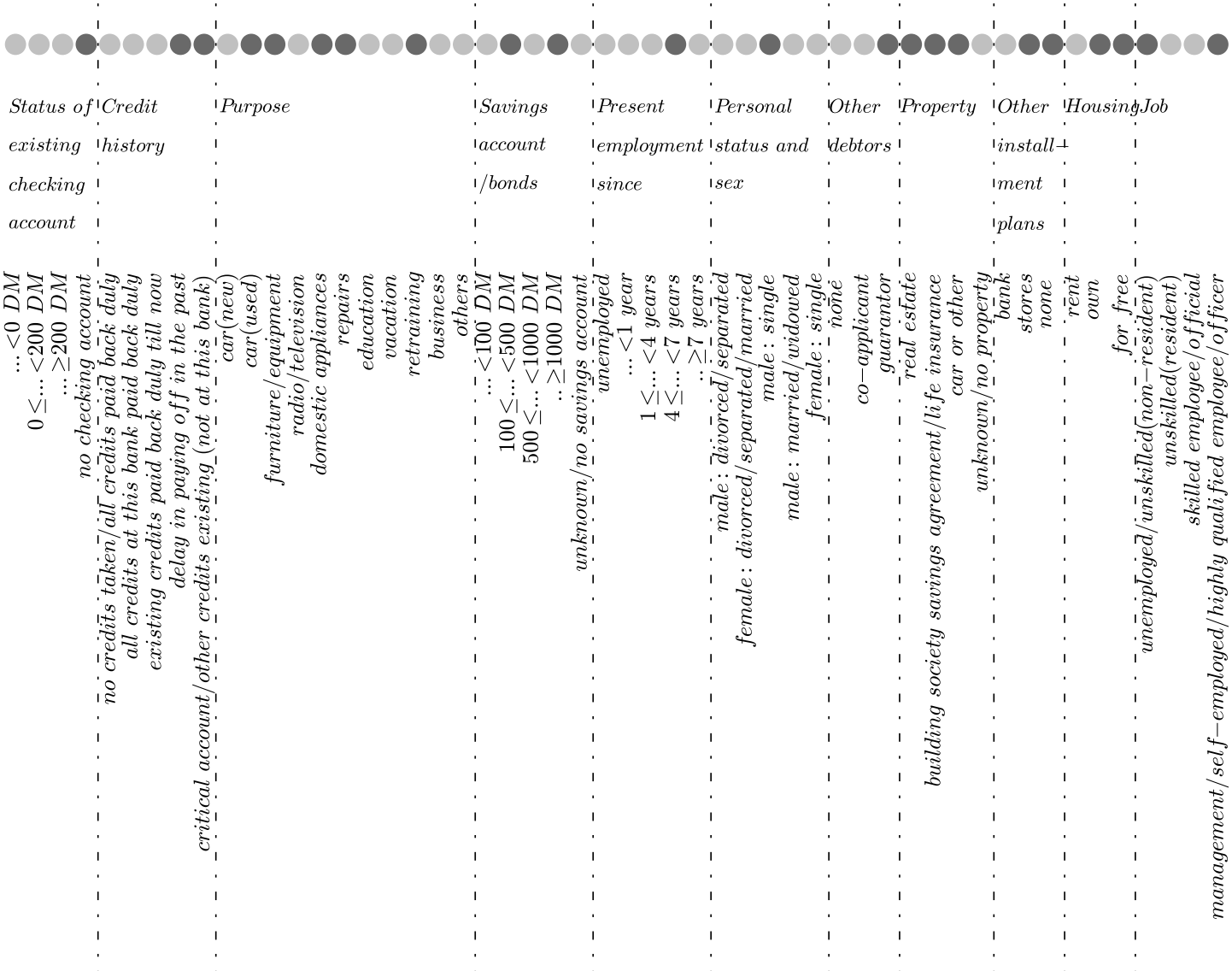
Given a categorical feature, if all its categories belong to the same cluster, then the feature is irrelevant to the CLSVM classifier.

Proposition

If $L_j = 2$, for a given j, for any optimal solution of CL, it holds that:

$$\bar{\omega}_{j,1} \cdot \bar{\omega}_{j,2} \leq 0.$$

An example for $L_j = 2$



A Mixed Integer Quadratic Programming (MIQP) formulation

$$\min_{\bar{\omega},\tilde{\omega},\omega',b,\xi,z} \sum_{j=1}^{J}\sum_{\ell=1}^{L_j} \frac{(\bar{\omega}_{j,\ell})^2}{2} + \sum_{j'=1}^{J'} \frac{(\omega'_{j'})^2}{2} + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

s.t. (CL-bigM)

$$y_i \left(\sum_{j=1}^{J}\sum_{\ell=1}^{L_j} \tilde{\omega}_{j,k(i),\ell} + (\omega')^\top x'_i + b \right) \geq 1 - \xi_i \qquad i = 1,\dots,n$$

$$\sum_{\ell=1}^{L_j} z_{j,k,\ell} = 1 \qquad k = 1,\dots,K_j, \quad j = 1,\dots,J$$

$$\tilde{\omega}_{j,k,\ell} \leq \bar{\omega}_{j,\ell} + M(1 - z_{j,k,\ell}) \qquad k = 1,\dots,K_j, \quad \ell = 1,\dots,L_j, \quad j = 1,\dots,J$$

$$\tilde{\omega}_{j,k,\ell} \geq \bar{\omega}_{j,\ell} - M(1 - z_{j,k,\ell}) \qquad k = 1,\dots,K_j, \quad \ell = 1,\dots,L_j, \quad j = 1,\dots,J$$

$$\tilde{\omega}_{j,k,\ell} \leq M\, z_{j,k,\ell} \qquad k = 1,\dots,K_j, \quad \ell = 1,\dots,L_j, \quad j = 1,\dots,J$$

$$\tilde{\omega}_{j,k,\ell} \geq -M\, z_{j,k,\ell} \qquad k = 1,\dots,K_j, \quad \ell = 1,\dots,L_j, \quad j = 1,\dots,J$$

$$\xi_i \geq 0 \qquad i = 1,\dots,n$$

$$z \in \{0,1\}^{\sum_{j=1}^{J} L_j K_j}$$

$$\bar{\omega} \in \mathbb{R}^{\sum_{j=1}^{J} L_j}$$

$$\omega' \in \mathbb{R}^{J'}$$

$$\tilde{\omega} \in \mathbb{R}^{\sum_{j=1}^{J} L_j K_j}$$

$$b \in \mathbb{R}$$

An MIQP formulation

$$\min_{\bar{\omega},\tilde{\omega},\omega',b,\xi,z} \sum_{j=1}^{J}\sum_{\ell=1}^{L_j} \frac{(\bar{\omega}_{j,\ell})^2}{2} + \sum_{j'=1}^{J'} \frac{(\omega'_{j'})^2}{2} + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

s.t. (CL-bigM)

$$y_i \left(\sum_{j=1}^{J}\sum_{\ell=1}^{L_j} \tilde{\omega}_{j,k(i),\ell} + (\omega')^\top x'_i + b \right) \geq 1 - \xi_i \qquad i = 1, \ldots, n$$

An MIQP formulation

$$\sum_{\ell=1}^{L_j} z_{j,k,\ell} = 1 \qquad k = 1, \ldots, K_j, \;\; j = 1, \ldots, J$$

$$\tilde{\omega}_{j,k,\ell} \leq \bar{\omega}_{j,\ell} + M(1 - z_{j,k,\ell}) \qquad k = 1, \ldots, K_j, \;\; \ell = 1, \ldots, L_j, \;\; j = 1, \ldots, J$$

$$\tilde{\omega}_{j,k,\ell} \geq \bar{\omega}_{j,\ell} - M(1 - z_{j,k,\ell}) \qquad k = 1, \ldots, K_j, \;\; \ell = 1, \ldots, L_j, \;\; j = 1, \ldots, J$$

$$\tilde{\omega}_{j,k,\ell} \leq M\, z_{j,k,\ell} \qquad k = 1, \ldots, K_j, \;\; \ell = 1, \ldots, L_j, \;\; j = 1, \ldots, J$$

$$\tilde{\omega}_{j,k,\ell} \geq -M\, z_{j,k,\ell} \qquad k = 1, \ldots, K_j, \;\; \ell = 1, \ldots, L_j, \;\; j = 1, \ldots, J$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, n$$

$$z \in \{0,1\}^{\sum_{j=1}^{J} L_j \, K_j}$$

$$\bar{\omega} \in \mathbb{R}^{\sum_{j=1}^{J} L_j}$$

$$\omega' \in \mathbb{R}^{J'}$$

$$\tilde{\omega} \in \mathbb{R}^{\sum_{j=1}^{J} L_j \, K_j}$$

$$b \in \mathbb{R}$$

Strategies to build the CLSVM classifier

- SVM$^{\mathrm{C}}$ Strategy:
  - solve the SVM
  - cluster the SVM scores$\rightarrow$ assignment vector z^*
  - solve the SVM for the clustered dataset

- CL$^{\mathrm{RR}}$ Strategy:
  - solve the relaxation of CL$\rightarrow$ fractional vector z
  - apply a randomized rounding procedure to $z \rightarrow z^*$
  - solve the SVM for the clustered dataset

- CLM$^{\mathrm{RR}}$ Strategy: similar as CL$^{\mathrm{RR}}$ Strategy for CLM

- CLM Strategy: obtain the CLSVM classifier directly from CLM

Computational results

Benchmark the CLSVM classifier against the SVM

In terms of

- Classification accuracy
- Complexity (for SVM: $\frac{card(\{\omega_{j,k} \neq 0\})}{\sum_{j=1}^{J} K_j}$, for CLSVM: $\frac{card(\{\bar{\omega}_{j,\ell} \neq 0\})}{\sum_{j=1}^{J} K_j}$)

The datasets

Name	$\lvert\Omega\rvert$	n	Class split (in %)	J	J'	$\sum_{j=1}^{J} K_j$	K_j
census income	95130	5000	94/6	31	9	491	9,52,47,17,3,7,24,15,5,10,3,6,8,6,6, 50,38,8,9,8,9,3,3,5,42,42,42,5,3,3,3
adult	30956	5000	24/76	11	3	117	5,8,5,16,5,7,14,6,5,5,41
mushrooms	8124	5000	48/52	17	4	111	6,4,10,9,4,3,12,4,4,9,9,4,3,8,9,6,7
coil 2000	5822	3900	94/6	5	80	77	41,6,10,10,10
abalone	4177	2800	50/50	1	7	3	3
molecular	3190	2200	52/48	60	0	480	8,8,8,. . .
careval	1728	1200	30/70	6	0	21	4,4,4,3,3,3
solar-c	1066	800	83/17	5	5	23	7,6,4,3,3
german	1000	700	30/70	11	9	52	4,5,11,5,5,5,3,4,3,3,4
australian	690	500	56/44	4	10	29	3,14,9,3

Setup

- $\frac{C}{n} \in \{10^{-6}, \ldots, 10^{6}\}$
- $M = 1000$
- time limit of 300 seconds for CLM Strategy

- Ipopt & Neos Server for the CL$^{\mathrm{RR}}$
- CPLEX v12.5 for the SVM, SVM$^{\mathrm{C}}$, CLM and CLM$^{\mathrm{RR}}$

Accuracy results for the SVM and the CLSVM methodology

$Name$	SVM		
	mean	std	med
census income	94.90	0.00	94.90
adult	84.57	0.22	84.63
mushrooms	100.00	0.00	100.00
coil 2000	100.00	0.00	100.00
abalone	79.87	1.18	79.72
molecular	94.22	0.80	94.04
careval	96.74	1.34	96.97
solar-c	83.53	1.23	83.46
german	74.60	2.71	75.66
australian	84.11	3.17	84.73

$Name$	SVM^{C}			CL^{RR}			CLM^{RR}			CLM		
	mean	std	med	mean	std	med	mean	std	med	mean	std	med
census income	94.85	0.00	94.85	94.84	0.04	94.82	94.40	0.04	94.37	94.37	0.00	94.37
adult	88.22	2.44	89.59	83.44	0.37	83.44	88.75	2.96	89.63	85.35	3.16	83.44
mushrooms	100.00	0.00	100.00	100.00	0.00	100.00	98.58	0.77	98.59	100.00	0.00	100.00
coil 2000	100.00	0.00	100.00	100.00	0.00	100.00	100.00	0.00	100.00	100.00	0.00	100.00
abalone	79.90	1.05	79.44	79.86	1.02	79.44	79.87	0.96	79.66	79.65	1.25	79.29
molecular	93.94	0.73	93.84	93.40	1.28	93.53	88.04	1.68	88.38	51.92	0.00	51.92
careval	82.80	5.15	82.95	92.23	1.28	92.42	83.94	4.91	85.41	94.17	2.84	93.37
solar-c	83.61	1.38	83.46	83.83	1.08	83.46	83.76	1.02	83.46	83.61	1.38	83.46
german	74.80	2.36	75.00	74.60	3.12	74.00	72.53	3.77	71.66	75.60	3.01	76.34
australian	84.42	3.32	84.73	84.53	3.12	84.73	84.53	3.05	84.73	85.37	3.28	85.26

Complexity results for the SVM and the CLSVM methodology

$Name$	SVM		
	mean	std	med
census income	63.14	0.00	63.14
adult	83.93	3.26	84.62
mushrooms	71.17	0.00	71.17
coil 2000	1.30	0.00	1.30
abalone	100.00	0.00	100.00
molecular	57.04	3.12	58.54
careval	99.05	2.86	100.00
solar-c	52.17	34.51	69.57
german	94.62	1.88	95.19
australian	86.90	4.83	89.66

$Name$	SVM^{C}			CL^{RR}			CLM^{RR}			CLM		
	mean	std	med	mean	std	med	mean	std	med	mean	std	med
census income	10.18	0.00	10.18	8.31	0.33	8.55	8.55	1.62	8.35	0.00	0.00	0.00
adult	13.76	2.21	13.68	16.58	1.34	16.24	10.17	4.36	8.55	9.23	3.64	9.83
mushrooms	23.42	0.00	23.42	19.28	4.56	18.02	21.71	5.87	22.07	14.77	1.57	14.41
coil 2000	1.30	0.00	1.30	1.30	0.00	1.30	1.30	0.00	1.30	1.30	0.00	1.30
abalone	66.67	0.00	66.67	66.67	0.00	66.67	66.67	0.00	66.67	66.67	0.00	66.67
molecular	0.00	0.00	0.00	24.87	0.19	25.00	22.71	0.90	22.50	0.00	0.00	0.00
careval	44.76	8.57	38.10	41.90	12.2	38.10	29.52	10.82	28.57	49.52	3.81	47.62
solar-c	12.17	10.43	8.70	5.22	11.14	0.00	0.87	2.61	0.00	11.31	7.83	8.70
german	42.31	0.00	42.31	38.27	4.59	41.35	36.92	4.28	36.54	42.31	0.00	42.31
australian	19.31	9.66	24.14	15.17	11.03	13.79	5.52	2.76	6.90	23.45	5.52	27.59

Summary

We propose the CLSVM methodology:

- Letting categories cluster around their peers
- Building an SVM-type classifier
- Comparable accuracy to the SVM
- Dramatic improvement in complexity

Future research

- Knowledge domain
- Sequential methodology for larger datasets
- Extend to continuous features:
    - Discretizing
    - Binarizing